POINT PROCESS MODELS OF SPECTRO-TEMPORAL MODULATION EVENTS FOR SPEECH RECOGNITION

Aren Jansen[1,2], Nima Mesgarani[2,3], Partha Niyogi[4]

[1]Human Language Technology Center of Excellence, [2]Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21211
[3]Department of Neurological Surgery, University of California, San Francisco, CA 94143
[4]Departments of Computer Science and Statistics, University of Chicago, Chicago, IL 60637
aren@jhu.edu, Nima.Mesgarani@ucsf.edu, niyogi@cs.uchicago.edu

ABSTRACT

Neurobiological research has uncovered the existence of cortical neurons in various animal species tuned to particular spectro-temporal modulations (STM) in the auditory stimulus. Other findings indicate that temporal statistics of the resulting neural spike trains may encode the underlying content of species-specific communication calls. With this motivation, we present an alternative approach to speech recognition based on point process statistical models of the local maxima events produced by a cortically-inspired spectro-temporal filter bank. We demonstrate the computational adequacy of this approach on the practical task of keyword spotting.

Index Terms— point process models, spectro-temporal modulation features, speech recognition

1. INTRODUCTION

The auditory cortices of humans and animals contain an enormous diversity of neurons that are tuned selectively to a wide range of acoustic stimuli. Moreover, several studies have demonstrated that such complex cells can reliably produce as few as a single reliable spike in response to a complex auditory stimulus, indicating a clear role for temporal coding of auditory information [1, 2]. The spectro-temporal receptive fields (STRF) of individual cortical neurons can be strongly correlated with a particular acoustic component of the species-specific communication system [3, 4]. However, it is less clear whether those receptive fields are innate (e.g. one filter in a parameterizable filterbank) or stimulus-derived (e.g. through learning or attention [5]).

The computational mechanism of sparse, temporal coding of selective neural activity described in the above references has been recently translated to speech technology in the form of point process models (PPM) for broad class/phonetic recognition [6, 7], keyword spotting [8], and continuous word recognition [9]. The PPM architecture is remarkably simple: (i) transform the speech signal into sparse point patterns of acoustic events in time and (ii) decode linguistic objects (e.g. words, syllables, consonant clusters) with explicit models of the temporal statistics of these patterns. One major design challenge under this architecture is the choice of which salient features of the speech signal should trigger acoustic events for subsequent modeling. In previous work, distinctive features [6], phonetic features [8], and features tied to the states of hidden Markov models [9] have each been considered with reasonable success. However, in each of these cases the event detectors were built on top of heavily supervised, language-dependent classifiers.

With the success of whole word PPMs built on top of event streams produced with supervised methods, we are left with this question: does there exist a universal set of low level acoustic events whose temporal statistics would be sufficient to recognize words? Initial experimentation with common low-dimensional speech features (e.g. auditory spectrogram, mel-frequency cepstral coefficients) was met with limited success. The problem was rooted in the strong speaker dependence of such representations; combining word-level acoustics across several speakers into a flat PPM led to a destruction of phonetic discriminability beyond broad classes.

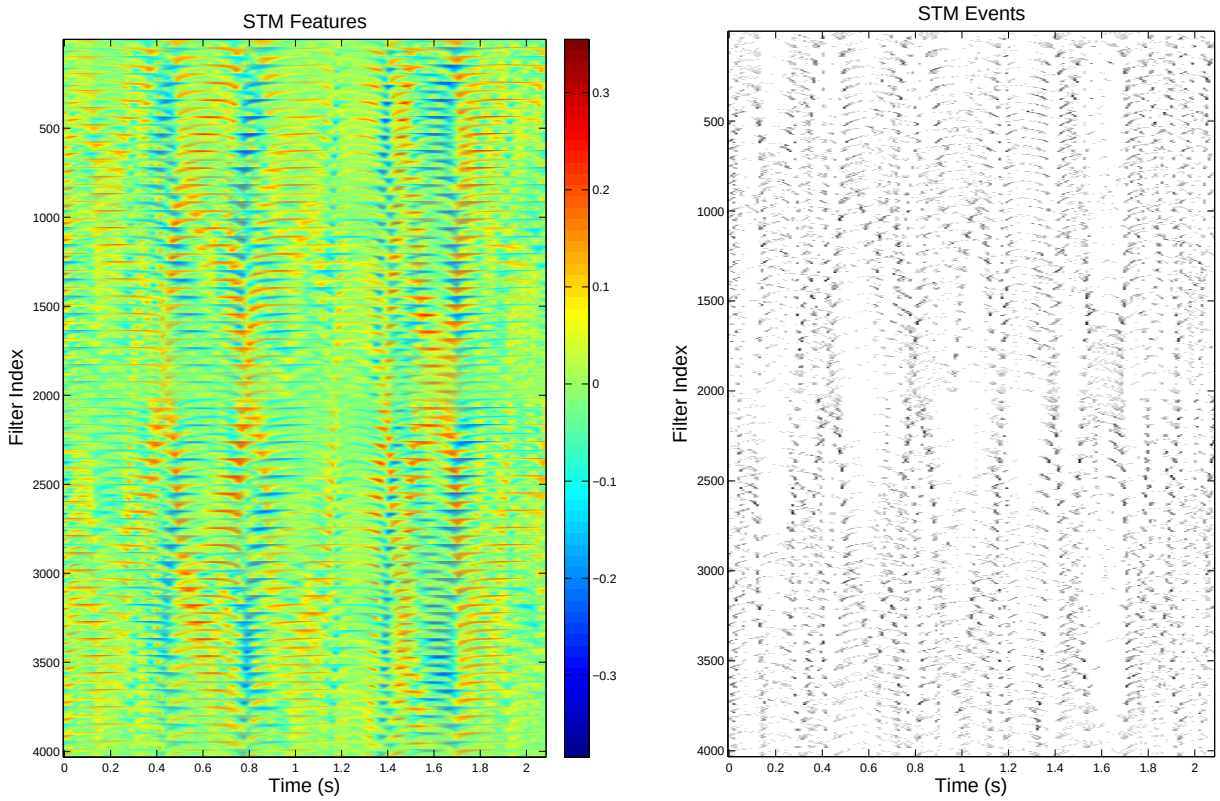
Overcomplete, multiresolution spectro-temporal wavelet decompositions, where each component corresponds to a oriented modulation in the time-frequency plane, provide a richer feature set while maintaining independence from any phonological system. Moreover, there is evidence that the individual components are similarly tuned as real cortical STRFs measured in animals [10]. While such spectro-temporal modulation features are not phonologically motivated, each individual filter can exhibit strong correlation with individual phones or small phonetic categories [10], performing a sort of universal phonological decomposition.

The class of spectro-temporal modulation (STM) features have been recently evaluated for speech applications including speech/non-speech discrimination [11], phonetic recognition [12], keyword spotting [13], and word recognition [14]. However, in each case a frame-based strategy was employed, which ultimately relied on dimensionality reduction to improve computational tractability. The point process model framework provides an alternative recognition pipeline that can leverage temporal sparsity to accommodate high dimensional STM features without subspace projections.

In this paper, we apply whole word point process models to sparse STM events. By building word models directly on this language independent and speaker dependent representation, we shift the entire acoustic modeling burden to the PPM. We show that the PPM is indeed capable of shouldering it, producing keyword spotting performance that is comparable to two systems based on supervised English phonetic acoustic models. In conjunction with detection-based ASR strategies [9], this result has implications for continuous word recognition more generally.

2. SPECTRO-TEMPORAL MODULATION EVENTS

Our goal is to define a family of acoustic event detectors $\{D_\phi\}_{\phi \in \Phi}$ for some acoustic feature set Φ that take as input a speech signal and each produce a sparse set N_ϕ of activation points in time (see Figure 1). Each detector D_ϕ is the composition of two operations:



(a) (b)

Fig. 2. An example (a) STM filterbank output and (b) the corresponding point process representation ($\delta = 0.05$) for the utterance "a mandatory retirement age of seventy."

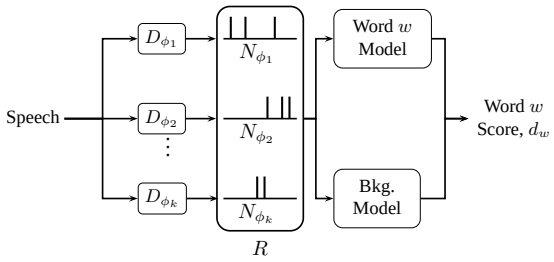


Fig. 1. High-level model architecture.

(i) the application of a feature dependent function g_ϕ to the waveform that takes high values only when feature ϕ is present, and (ii) the application of a thresholded peak finding function to the output of g_ϕ that computes the event set N_ϕ. In this study, we choose our acoustic feature set Φ to be an overcomplete set of spectro-temporal modulations. These modulations may be detected in the speech signal by applying a family of 2-dimensional Gabor filters, which are parameterized by the following three quantities (see [11] for a comprehensive treatment):

- Center *frequency* of the modulation
- Temporal modulation *rate* (temporal resolution) in Hz
- Frequency modulation *scale* (frequency resolution) in cycles/octave

We consider a 3-dimensional grid of STM filter parameters that includes all possible combinations of $\mathcal{F} = 21$ center frequencies (the first 21 Bark bands up to the Nyquist rate of 8 kHz), $\mathcal{R} = 16$ rates (16 logarithmically spaced values from -22.6 Hz to 22.6 Hz, where positive/negative values imply upward/downward selectivity), and $\mathcal{S} = 6$ scales (logarithmically spaced from 1/16 to 1 cycles/octave). Thus, our feature set Φ is defined such that each feature ϕ corresponds to a particular STM filter with a response characterized by a single point in the parameter grid $|\Phi| = 2 \times \mathcal{F} \times \mathcal{R} \times \mathcal{S} = 4032$. Here, the extra factor of 2 arises from including both symmetric and asymmetric filters for each modulation (corresponding to the real and imaginary part of each complex-valued output, respectively). As such, the output of this STM filterbank defines a 4032-dimensional vector time series. An example is displayed in Figure 2(a) for the utterance "a mandatory retirement age of seventy." In this figure, the filter index is ordered by the expression $(c-1)\mathcal{F}\mathcal{R}\mathcal{S} + (f-1)\mathcal{R}\mathcal{S} + (r-1)\mathcal{S} + s$, where f, r, s are the parameter grid indices for frequency, rate, and scale, respectively, while $c = 1$ indicates symmetric and $c = 2$ asymmetric.

Each STM filter ϕ (corresponding to a single row in Figure 2) determines a scalar time series g_ϕ that takes high values when the given spectro-temporal modulation is present in the signal. We define our spectro-temporal modulation events for STM ϕ as the set N_ϕ of all times that g_ϕ experiences a local maxima above some predefined threshold δ:

$$N_\phi = \{i\Delta | g_\phi[i] > g_\phi[i \pm 1], g_\phi[i] \geq \delta\}, \quad (1)$$

where Δ is the sampling period of the time series. After the application of the threshold, no signal amplitude information is stored

or modeled, resting the subsequent recognition burden solely on the timings of the STM events. We collect all event sets into a single event pattern representation $R = \{N_\phi\}_{\phi\in\Phi}$ used for subsequent word modeling in Section 3. Figure 2(b) displays the composite event representation R produced from the STM filterbank output shown in Figure 2(a), where a threshold of $\delta = 0.05$ was applied. At this threshold, there are on average only 140 of the 4032 detectors activated at any given time step, resulting in a sort of dynamic dimensionality reduction.

3. POINT PROCESS WORD MODELS

Insofar as the STM events are analogous to selective neuronal firings, the representation of Figure 2(b) can be interpreted as a spike raster plot for 4032 neurons responding the given stimulus. The question is now whether such an event pattern is sufficient information to decode the underlying word sequence. With this motivation, our next task is to define a model for each target word w that facilitates mapping observed composite STM event patterns R to a word detector time series $d_w(t)$ (see Figure 1). For this, we turn to the point process model (PPM) framework, in the form originally proposed in [8]. The PPM assumes that two underlying stochastic processes produce each observed composite STM event pattern R. The first is a background homogeneous Poisson process that generates the event patterns not contained within any instances of a target word w. The second is an inhomogeneous Poisson process that generates the point pattern observed within instances of the target word w. The hidden variable that triggers a word occurrence (and thus the word process) is an indicator function of time $\theta_w(t)$, which takes a value 1 at the beginning of an occurrence of w and is 0 everywhere else.

Given this formalism, we can define the word w detector function, $d_w(t)$, as the log likelihood ratio

$$d_w(t) = \log\left[\frac{P(R|\theta_w(t)=1)}{P(R|\theta_w(t)=0)}\right], \quad (2)$$

If we let T be a latent random variable for the word duration, we can partition the observed STM event pattern R for an utterance of duration L into three subsets: $R_a = R|_{(0,t]}$, $R_{t,T} = R|_{(t,t+T]}$, and $R_b = R|_{(t+T,L]}$. Here $R_{t,T}$ are the observed events within the candidate word interval $(t, t+T]$, while R_a and R_b are the observations to the left and right, respectively. Since we will be assuming Poisson processes, we can assume conditional independence of the three subsets. Moreover, since R_a and R_b are outside the candidate word interval $(t, t+T]$, we assume they are generated by the same homogeneous background process. Thus, the likelihoods of R_a and R_b will cancel out in the ratio. Noting that $P(R|\theta_w(t)=0)$ does not depend on T and expanding the word likelihood to include the integral over all possible durations T, Equation 2 takes the form

$$d_w(t) = \log\left[\int_0^{L-t} \frac{P(R_{t,T}|T,\theta_w(t)=1)}{P(R_{t,T}|T,\theta_w(t)=0)} P(T|\theta_w(t)=1)dT\right]. \quad (3)$$

The duration likelihood $P(T|\theta_w(t)=1)$ can either be assumed uniform with bounded support over some interval $[T_{\min}, T_{\max}]$, or be estimated from a set of word examples using kernel density estimation. Next, $P(R_{t,T}|T,\theta_w(t)=0)$ and $P(R_{t,T}|T,\theta_w(t)=1)$ are determined by the background homogeneous Poisson and inhomogeneous Poisson word model, respectively, as follows:

1. For the $P(R_{t,T}|T,\theta_w(t)=1)$ distribution, we begin by normalizing $R_{t,T}$ to the interval $(0,1]$; that is, we map $R_{t,T}$ to $R'_{t,T}$ such that for each $t_i \in R_{t,T}$ there is a corresponding $t'_i \in R'_{t,T}$ where $t'_i = [t_i - (t-T)]/T$. Given this mapping, we make the simplifying assumption that

$$P(R_{t,T}|T,\theta_w(t)=1) = \frac{1}{T^{|R_{t,T}|}} P(R'_{t,T}|\theta_w(t)=1). \quad (4)$$

This equivalence assumes that the observations for each instance of the word are generated by a common, T-independent inhomogeneous Poisson process operating on the interval $(0,1]$ that is subsequently scaled by T to a point pattern on the interval $(t, t+T]$. In this way, we assume the number of firings of the different STM event detectors present in a word is invariant to the actual duration of the word. Assuming an inhomogeneous Poisson process generated $R'_{t,T}$, its likelihood takes the form

$$P(R'_{t,T}|\theta_w(t)=1) = \prod_{\phi\in\Phi} e^{-\int_0^1 \lambda_\phi(s)ds} \prod_{s\in N'_\phi} \lambda_\phi(s), \quad (5)$$

where $\lambda_\phi(s)$ is the rate parameter at normalized time $s \in (0,1]$ for STM ϕ and N'_ϕ are the elements of $R'_{t,T}$.

Learning a new word amounts to learning the rate parameter functions $\{\lambda_\phi\}_{\phi\in\Phi}$ from STM events observed when presented with examples of the word. This can be accomplished with kernel smoothing, but in this study we consider instead a maximum likelihood estimate using a piecewise-constant parametric model of each λ_ϕ with D uniformly delimited segments. Figure 3 displays the word model rate parameters for the word "Boston," estimated with training examples from three separate speakers. Even though the context, speaker rate, and gender varies across the examples, their consolidation into a single word model produces rate parameters with a distinct temporal structure.

2. For $P(R_{t,T}|T,\theta_w(t)=0)$, for which we consider a homogeneous Poisson process model, the likelihood depends solely on the total number n_ϕ of events observed for each STM ϕ and the total duration of the segment (in this case T). The likelihood given this homogeneous Poisson process model takes the form

$$P(R_{t,T}|T,\theta_w(t)=0) = \prod_{\phi\in\Phi} [\mu_\phi]^{n_\phi} e^{-\mu_\phi T}, \quad (6)$$

where μ_ϕ is the background rate parameter for STM ϕ and n_ϕ are the number of elements in $R_{t,T}$ for STM ϕ. Training this model requires estimating the rate parameters as the average detector firing rates over a collection of arbitrary speech.

Given a novel utterance, we may evaluate the detector function by sliding a set of windows with durations $T \in \mathcal{T}$ and approximating the integral expression of Equation 3 with an appropriately weighted sum.

4. EXPERIMENTS

We borrowed the experimental protocol from [8], a keyword spotting evaluation using the Boston University Radio News (BURadio) Corpus. BURadio consists of over 7 hours of 16 kHz/16 bit newsreader style recordings of 7 speakers (4 males and 3 females). We

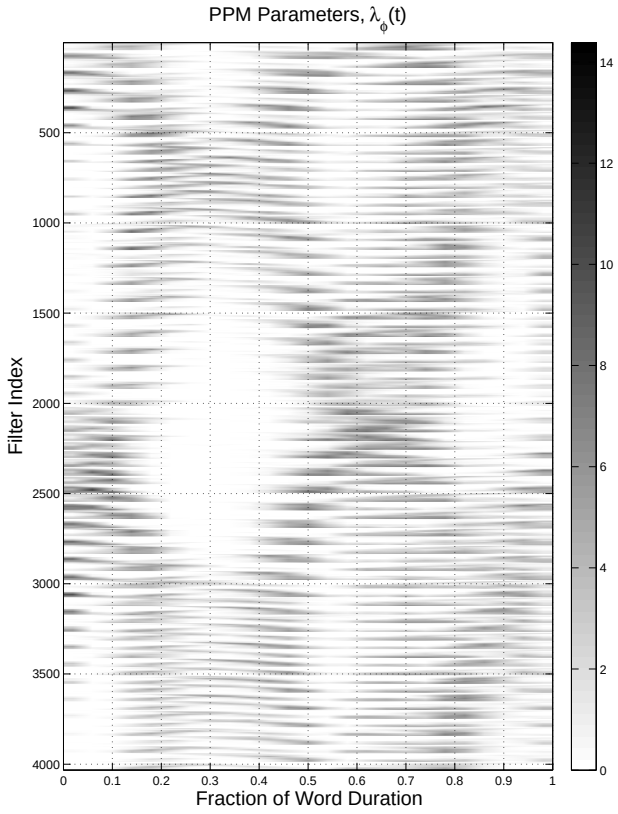


Fig. 3. Word model Poisson rate parameters, $\{\lambda_\phi(t)\}_{\phi\in\Phi}$, for the word "Boston," where we have set $D = 20$.

partitioned the speakers into a training set of two females and two males (f1a,f3a,m1b,m2b) and a test set of the remaining speakers (f2b,m3b,m4b). The first column of Table 1 lists the evaluation set of 13 keywords (11 content, 2 function) pruned from the original set of 20 by removing those that had fewer than 50 training examples.[1]

Our STM event PPM-based keyword spotter was built on top of 4032-dimensional STM features derived from an auditory spectrogram using a 25 ms Hamming window and 10 ms step. Each whole word PPM was trained using all training set word examples to estimate piecewise constant estimates of the Poisson rate functions with $D = 10$. The background rate constants $\{\mu_\phi\}_{\phi\in\Phi}$ were estimated using a random sample of 30 minutes of training speech.

4.1. Baselines

We used the systems evaluated in [8] to provide two keyword spotting baselines. The first was a keyword-filler hidden Markov model-based system with an architecture similar to [15], built on top of a Gaussian mixture model (GMM) based monophone acoustic model (one 8 component GMM per phone with a standard 39-dimensional mel frequency cepstral coefficient observation space, trained on the TIMIT sx/i sentences). Each keyword HMM consisted of one pronunciation path per word with the phone GMMs defining the emission densities. The filler model consisted of a phone loop with the transition probabilities estimated from BURadio training data. The second baseline was a phonetic event point process model, which used an identical architecture to the present system, but with a phone feature set replacing the present STM features. The phone PPM system relied on phonetic event detectors constructed with the same GMM-based monophone model used in the keyword-filler HMM. Complete details of both baseline systems can be found in [8].

[1]The present models require $\sim$ 50 examples for successful estimation; see the end of Section 4.2 for a discussion.

4.2. Results and Discussion

Following the evaluation prescription defined in [8], we train each word model from all instances present in the training set and test each word detector on at least 1 hour of test speech that contains all test instances of each word. For each of the 13 words in the evaluation set, Table 1 lists the number of training examples, median word example duration (from training set), and the performance of each of the three systems (STM event PPM, phonetic event PPM baseline, and keyword-filler HMM baseline). The performance figure-of-merit is defined as the mean detection rate over the range of operating points that produce 1-10 false alarms per hour (i.e. the area under the initial portion of the ROC curve, roughly the detection rate at 5 false alarms per hour). There are several properties of these results to note:

1. In prior studies, it has been demonstrated that despite having equivalent average performance, PPMs based on phonetic events have significantly different error patterns than equivalent frame-based HMMs, even though they are built on top of the same GMM-based phonetic acoustic model. This indicates a strong complementarity of the two approaches that may provide further gains with system combination. Here, we find that the STM event PPMs share a comparable average performance across the word set, but again exhibit a significantly different error pattern compared to both the phonetic PPM and HMM systems. Thus, we conclude that the STM events are producing information that is distinct and complementary to the phonetic information produced by the GMMs.
2. While the average performance of the three systems are equivalent, it is of paramount importance to emphasize that the STM event PPMs are accomplishing this with no background knowledge of the acoustic-phonetics of English. That is, the STM filters, which are tied to the properties of the auditory cortex but not any particular language, provide local maxima events whose timings are sufficient to recognize words as well as a phonologically organized and supervised acoustic model. That the STMs are not tied to the phone set of English means we can expect similar word spotting performance in any language provided we have a sufficient number of examples to train the PPM.
3. In general, the point process modeling framework can support events marked with a measure of their confidence (see [7]). However, in the present system we have discarded all STM filterbank output amplitudes and decode based on the timing of the STM events alone. The relative success of this strategy provides evidence for the computational power of sparse temporal coding of action potentials, which is thought to play a significant role in cognitive function [2].
4. It is important to emphasize that the STM filters have no knowledge of speaker variability, whereas the phonetic GMMs have been presented with large amount of speakers from the TIMIT database. Still the STM event PPMs are able to generalize its knowledge of each word provided from 4 training speakers to a completely distinct set of 3 speakers at test time.

Table 1. Keyword spotting figure-of-merit results for the three systems.

Keyword	# Train	Median T	STM PPM	Phone PPM	HMM
Massachusetts	334	710 ms	94.7	**98.5**	98.0
Boston	272	470 ms	**91.8**	89.3	85.7
president	52	490 ms	82.3	**83.0**	71.8
percent	80	450 ms	**79.3**	71.3	65.9
public	68	340 ms	**75.8**	60.1	60.6
thousand	56	490 ms	73.0	78.9	**85.7**
yesterday	90	550 ms	57.5	59.6	**89.0**
hundred	121	310 ms	33.5	34.1	**49.5**
state	273	300 ms	**32.4**	26.6	23.3
year	144	230 ms	30.7	**33.1**	20.3
about	116	250 ms	27.9	27.9	**30.2**
by	337	180 ms	**19.2**	9.0	9.8
time	82	320 ms	17.1	**25.7**	21.8
		Averages:	**55.1**	53.6	54.7

The one major drawback of the current implementation is the dependence on a fairly significant number of word training examples (~ 50) to estimate the large number of model parameters ($D|\Phi| \approx$ 80k). However, we believe that either kernel density estimation with the optimal Epanechnikov kernel or a more conservative parametric model (e.g. 1-D Gaussians) can be employed to reduce the word example requirement and we will investigate that in future work. Moreover, discriminative model training that incorporate appropriate regularizers might also further reduce the training requirements.

5. CONCLUSIONS

We have presented a new word modeling technique based on the temporal statistics of events triggered by spectro-temporal modulations present in the speech signal. The point process modeling framework provides an efficient way to use high dimensional spectro-temporal modulation features. By considering a word spotting task, we demonstrated that sparse temporal patterns of spectro-temporal modulation events are sufficient to recognize words with accuracy comparable to techniques dependent on a supervised phonetic acoustic model. Thus, the proposed technique facilitates speaker-independent keyword spotting in any language given only as few as 50 examples of a word.

6. REFERENCES

[1] K.-H. Esser, C. J. Condon, N. Suga, and J. S. Kanwal, "Syntax processing by auditory cortical neurons in the FM-FM area of the mustached bat pteronotus parnellii," *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 14019–14024, 1997.

[2] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Curr. Opin. Neurobiol.*, vol. 14, pp. 481–487, 2004.

[3] K. Sen, F. E. Theunissen, and A. J. Doupe, "Feature analysis of natural sounds in the songbird auditory forebrain," *J. Neurophysiology*, vol. 86, 2001.

[4] N. Suga, "Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception," in *Listening to Speech: An Auditory Perspective (S. Greenberg and W. A. Ainsworth, Eds.)*, pp. 159–182. Lawrence Erlbaum Associcates, Mahwah, NJ, 2006.

[5] J. B. Fritz, S. V. David, S. Radtke-Schuller, P. Yin, and S. A. Shamma, "Adaptive, behaviorally gated, persisten encoding of task-relevant auditory information in ferret frontal cortex," *Nature Neuroscience*, vol. 13, no. 8, pp. 1011–1019, 2010.

[6] A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," *J. Acoust. Soc. Am.*, vol. 124, pp. 1739–1758, 2008.

[7] A. Jansen and P. Niyogi, "Point process models for event-based speech recognition," *Speech Communication*, vol. 51, pp. 1155–1168, 2009.

[8] A. Jansen and P. Niyogi, "Point process models for spotting keywords in continuous speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 1457–1470, 2009.

[9] A. Jansen and P. Niyogi, "Detection-based speech recognition with sparse point process models," in *Proc. of ICASSP*, 2010.

[10] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *J. Acoust. Soc. of Am.*, vol. 123, pp. 899–909, 2008.

[11] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. on Audio Speech Lang. Process.*, vol. 14, 2006.

[12] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *Proc. of ICASSP*, 2008.

[13] T. Ezzat and T. Poggio, "Discriminative word-spotting using ordered spectro-temporal patch features," in *Proc. of Interspeech*, 2008.

[14] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication (In Press)*, 2010.

[15] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát, and J. Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," in *Lecture Notes in Computer Science - TSD 2005 (V. Matousek et al.)*, pp. 302–309. Springer-Verlag, Berlin, 2005.